ChinaXiv合作期刊

基于主题模型的科技报告文档聚类方法研究*

曲靖野[1,2] 陈震[1] 郑彦宁[2]

[1] 北华大学信息技术与传媒学院 吉林 132013 [2] 中国科学技术信息研究所 北京 100038

摘要：[**目的/意义**]探索实践以科技报告为文献载体形式的融合主题模型的文本聚类方法，拓展基于科技文献进行技术监测服务的新领域，提出基于科技报告进行语义分析的新方法。[**方法/过程**]以国家科技报告服务系统中的科技报告为数据源，首先基于 LDA 主题模型对经过文本预处理的科技报告进行主题挖掘，再基于 Ward 与 K-means 相结合的聚类算法对包含主题分布信息的文本向量进行聚类分析，尝试提出一种适合科技报告文档聚类的文本挖掘新方法。[**结果/结论**]实验结果表明，LDA 主题模型能有效准确挖掘科技报告中的主题信息，所提出的 Ward 与 K-means 相结合的聚类算法对科技报告的聚类效果也优于其它传统聚类算法。

关键词：科技报告 主题模型 LDA 文本聚类

分类号：G203

DOI：10.13266/j.issn.0252-3116.2018.04.015

chinaXiv:202308.00385v1

1 引言

科技报告作为科技文献的重要组成部分，是国家基础性、战略性科技资源，是国家科技实力的重要体现。为深入实施创新驱动发展战略，2013 年科技部组织开展了科技部主管的国家科技计划项目科技报告制度建设工作，标志着国家对科技报告文献资源的发展和建设工作愈发重视。于此同时，相比于传统科技期刊和专利文献，科技报告文献资源本身内容翔实、专深，包含技术原理、方法、工艺和过程，具有重要的学术价值和实用价值，尤其在工程学领域，科技报告能提供关于技术实现过程、方法原理的精准描述，在技术描述的全面性及科技报告提交的实效性上明显优于传统的科技期刊资源[1]。

我国对科技报告相关研究的进度与科技报告建设工作的开展较为一致，内容较多是关于科技报告的概念及其体系和制度的建设等理论层面的研究。其中，比较有代表性的研究有：侯人华等针对科技报告的制度体系与形成模式进行了研究[2]；郭学武等对开放科技报告服务体系给出了建议[3]；毛刚等基于情报学视角对科技报告相关研究进行了解读[4]。相较于国内，国外发达国家比较重视科技报告的应用价值，美国和欧洲都分别建立了自己的科技报告平台[5-9]，比如美国的政府报告文摘题录数据库 NTIS、欧洲灰色文献信息系统 OpenGrey 等。我国于 2014 年 3 月起正式开通运行国家科技报告服务系统（NSTRS），其截止到 2017 年年初，科技报告数量达到 8.2 万余份[10]。该平台可以利用科技报告文献资源向社会提供检索、浏览、原文传递等相关信息服务，公众可以系统检索到国家和地方科技项目的各类报告。由此可见，随着近些年科技报告数量的爆炸式攀升以及跨学科研究的不断交融，对科技报告的研究急需从现有的理论政策层面过渡到信息组织、知识挖掘的层面。因此，如何针对海量科技报告文本中关于最新的技术、原理、方法等科技知识进行语义分析成为现阶段科技报告建设工作中亟需解决的一个主要问题。

基于科技报告的聚类方法研究可以生成内容相似的科技报告文档群，在科技报告智能检索、相关主题推送、技术监测服务等领域提供更好的服务。笔者提出一种以科技报告为载体数据源，基于主题识别与聚类方法相融合的科技报告文档聚类方法。这种聚类方法以经过 LDA 主题模型处理后的科技报告文档-主题向量为数据源，可以深入到科技报告文档内部的语义层面，从主题的视角对科技报告文档进行聚类研究。

* 本文系吉林省教育科学"十三五"规划项目"项目教学法在高校基础计算机教学中的应用研究"（项目编号：GH170061）研究成果之一。

作者简介：曲靖野（ORCID：0000-0002-1715-1919），副教授，博士；陈震（ORCID：0000-0002-3522-5272），高级实验师，博士，通讯作者，E-mail：59975235@qq.com；郑彦宁（ORCID：0000-0003-3885-7459），研究馆员，博士生导师。

收稿日期：2017-08-12 修回日期：2017-11-13 本文起止页码：113-120 本文责任编辑：王善军

ChinaXiv合作期刊

我国科技报告的建设尚处于起步阶段,因此,基于我国科技报告平台服务系统,以现阶段我国科技报告文本为数据源,探讨其文本预处理、主题识别及文档聚类融合的相关研究,对促进现阶段科技报告的语义挖掘、推动科技成果的开放共享、转化应用及对科技报告资源的深度开发利用具有一定的实践价值。

2 相关理论

2.1 LDA 主题模型理论

LDA 主题模型早期起源于隐含语义分析(Latent Semantic Analysis, LSA)[11],之后有学者利用概率论与数理统计的知识对其进行改进,提出了概率隐含语义分析(Probabilistic Latent Semantic Analysis,PLSA)[12],作为 LDA 的雏形,PLSA 继承了 LSA 算法的优点,并且可以解决 LSA 中多义词的问题,但是其词与主题之间的分布是固定的,处理文档的方法受到局限。2003 年 D. Blei 等人基于贝叶斯估计提出了隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)[13],LDA 由于使用了先验分布,所以待估算的参数随之减少,其处理文档的方法更加灵活[14]。

LDA 主题模型是一种贝叶斯版本的 PLSA 模型,其利用贝叶斯估计词分布与主题分布两个未知参数[15-16],LDA 主题模型有三个结构层次:特征词层、主题层和文档层,其工作原理如图 1 所示:

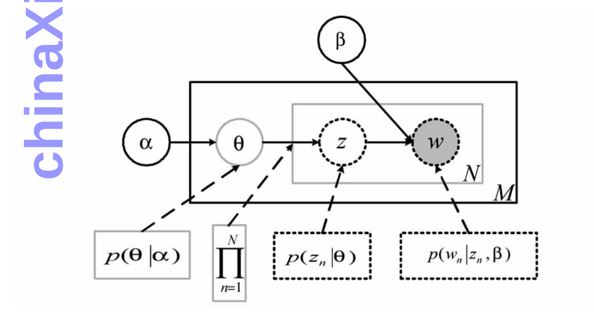


图 1 LDA 的原理模型

在图 1 中,灰色圆圈表示已知量,白色圆圈表示未知量,黑色实线代表语料层,灰色实线代表文档层,其中 θ 是一个主题向量,向量中的元素表示每一个主题在此文档中的出现概率。$p(\theta|\alpha)$ 为向量 θ 在给定参数 α 下的概率分布。黑色虚线代表特征词层,z 和 w 分别表示选取的主题和特征词,两者都是特征词级别的变量。$p(z_n|\theta)$ 为主题 z_n 在给定向量 θ 下的概率分布,$p(w_n|z_n,\beta)$ 为特征词 w_n 在给定主题 z_n 和参数 β 下的概率分布。LDA 主题层是模型的待求结果,其中 α 表示文档中主题的分布信息,β 表示主题中特征词的分布信息,通过这两者可以推算出作者感兴趣的主题以及每篇文章中的主题涵盖比例。

LDA 主题模型方法优势在于其具有严谨的概率统计理论基础作为支撑,可以通过以文档内容中的主题为单位,针对大批量样本以自动化的机器处理形式进行粒度更细致的信息提取和加工处理。传统的基于词频统计、引文分析、共词分析、内容分析等文献计量的主题识别方法[17-18]大多停留在基于科技文献外部特征的信息组织层面,词与主题的关联性及词之间的语义关系无法充分表征,无法针对数据样本进行语义分析。近年 LDA 主题模型被较多地应用到期刊、专利等传统科技文献的主题挖掘与演化等研究领域[19-22],如刘卫江在其硕士论文中以国外科技报告为例讨论了基于主题层面的科技监测方法[23]。笔者将基于 LDA 模型处理后所生成的文档 - 主题向量作为科技报告文档聚类的输入数据源,可以深入到科技报告文本的语义层面,从主题的视角对科技报告文本内容进行服务粒度更加细化的挖掘。

2.2 基于 Ward 与 K-means 相结合的文本聚类算法

由于科技报告文档数据量较大,经 LDA 模型对文档集进行处理后,所提取主题的数量也较多,并且不同科技报告文档可能出现主题分布概率相近所导致的研究内容相似等问题。因此,在对科技报告的主题进行挖掘处理之后,将具有相似主题的科技报告文本再次聚类,可以更好地在语义层面实现科技报告的检索、推送等知识服务。

传统的 K-means 算法是一种经典划分式聚类算法,其基本原理是通过自行选取 K_c 个文档(主题概率分布向量,下文简称为数据点)作为聚类的初始划分点,分别计算剩余数据点到 K_c 个划分点的距离,每个数据点与 K_c 个划分点中距离最近的划分为一类,通过计算每类中数据点的平均值来更新类中划分点的值,重复操作直到划分点稳定不变为止。可见 K-means 聚类算法的缺点是 K_c 的数目和初始数据划分点都要用户自行确定,而其初始参数设定的不合理和主观性很可能增加算法的时间复杂度,降低其聚类结果的精度。Ward 算法[24]能很好地弥补 K-means 算法的缺点,它是一种自下而上的凝聚层次式聚类算法,通过把每一个数据点作为一个初始类,把距离最近的两个类进行合并,合并后重新计算类间的距离[24],重复操作直到类的总数等于 1 或者满足预先设定的终止条件为止。Ward 算法能够自动确定分类的数量以及每个类的均值,这两个数值可以作为 K-means 算法中初始类的数

chinaXiv:202308.00385v1

量 K_c 和初始 K_c 个划分点的值,从而避免了 K-means 算法中人为设定这两个参数的主观性,提升了准确性及自动化程度。

因此,笔者基于 Ward 法与 K-means 法相结合的文本聚类算法,对经过 LDA 主题模型处理后的科技报告主题概率分布文档进行聚类分析,经实验可以更好地为科研人员提供有关该研究领域的相关研究主题的科技报告文档。

3 基于 LDA 模型的科技报告文本聚类方法设计

基于 LDA 模型的科技报告文本聚类方法处理流程主要为以下三个关键步骤:①科技报告数据的预处理;②基于 LDA 模型的科技报告主题识别方法设计;③文档聚类方法设计。具体如图 2 所示:

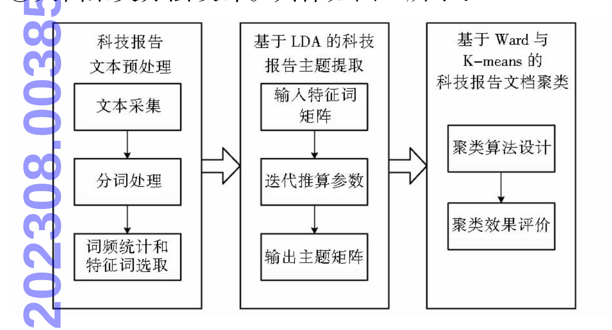


图 2 基于 LDA 模型的科技报告文本聚类方法设计

3.1 科技报告文本预处理

数据预处理是进行文本挖掘和主题识别的前期基础性工作,通过对科技报告原始文本数据的预处理过程,使其标准化、结构化从而适合作为机器自动化处理的 LDA 模型的输入,科技报告文本预处理结果的优劣直接决定着对其进行主题分析与文本聚类的结果优劣。科技报告文本数据预处理过程如图 2 所示,主要包括三部分:对科技报告文本数据的采集、对采集后的文本数据进行分词处理、对分词结果进行词频统计并最终保留最具标识性的文本特征词。

3.1.1 科技报告数据采集 科技报告是科研工作者按照规定的标准格式记录其从事的调查、研究、设计、实验和分析等科研工作的特殊文献。由于科技报告的内容专业、深入、详细,并且附有图表、研究方案、实验数据,所以能有效地体现该科研活动的各种信息,我国科技报告的组成见表 1。

表 1 我国科技报告的组成

	构成部分	作用	状态
前置部分	封面	提供描述性元数据信息	可选
	封二	提供权限等管理元数据	可选
	题名页	提供描述性信息	必备
	辑要页	提供描述及管理元数据	可选
	前言	提供描述元数据	可选
	摘要	提供描述元数据信息	必备
	关键词	提供结构元数据	必备
	目次	提供结构元数据	必备
	插图附表清单	提供结构元数据	可选
	符号说明	提供结构元数据	可选
正文部分	引言部分	内容	必备
	主体部分	内容	必备
	结论部分	内容	必备
	建议部分	内容	可选
	参考文献	结构元数据	必备
后置部分	附录	结构元数据	必备
	索引	结构元数据	可选
	发行列表	管理元数据	可选
	封底	提供描述元数据信息	可选

科技报告的说明信息主要分布在其前置部分,其中的核心是科技报告的摘要部分。科技报告的摘要作为其内容提要,是科技报告的重要构成,是报告全文的高度提炼。由于目前科技报告正文文本数据公开权限的限制,笔者选择采集科技报告摘要部分作为 LDA 主题模型所需的语料库输入,通过前期实验证实以此作为输入语料,主题挖掘效果较好,主题的语义边界较清晰。

3.1.2 科技报告文本分词处理 通过对科技报告摘要的文本采集所构建的语料库不能直接作为 LDA 模型的输入数据,其中的无效干扰词汇过多、维度过高,增加了计算成本,影响输出结果的准确性。因此,首先需要对初始语料库进行分词处理,分词处理包括两部分内容:①将连续的汉字序列通过分词算法切割成单独的词;②根据停用词表去掉数量词、副词、介词、连词、助词等干扰词汇。分词处理后得到的数据中每一行的词分别对应一篇摘要文档,数据的行数等于科技报告文档的篇数,分词处理就是把文档集按照一定的规则简化成数行词袋化的词向量的过程。目前中文分词方法可分为:理解分词法、词典分词法和统计分词法[25]。其中,利用统计分词算法设计的分词系统比较常用,具有代表性的工具有:支持多种编程语言的 Jieba 中文分词库、汉语词法分析系统(NLPIR/ICTCLAS)、在汉语词法分析系统的基础上通过优化算法和数据结构而编写的 Ansj 中文分词器。由于 Jieba 中文分词库的开源性和灵活性,笔者采用其对科技报告原始语料库进行分词。同时结合科技报告的特点,在

chinaXiv:202308.00385v1

ChinaXiv合作期刊

中文常用停用词表中加入“技术”“研究”“算法”等词汇以进行去停用词处理。

3.1.3 科技报告词频统计和特征词选取 经过分词处理后的原始语料的维度得到了初步的降低，但词向量结果集中的无效词仍然较多，需要进一步通过词频统计及特征词的选取实现主题模型的标准输入。词频统计是通过对某一词语在一定范围的文档中出现次数进行统计、整理、分析而得出关于该词语出现概率、分布范围等规律的一种统计方法。设定无效词具备以下特征：词频统计结果高于某一指定较高阈值的意义空虚的无效词；词频统计结果低于某一指定较低阈值的具有“长尾特征”的无效词。通过词频统计处理后所选取的特征词构成了表征科技报告文本数据主要特征的向量，这些向量构成了文档－特征词矩阵，作为 LDA 主题模型的有效数据输入。

3.2 基于 LDA 模型的科技报告主题识别方法设计

将经过科技报告文本数据预处理所得到的“文档－特征词矩阵”作为 LDA 主题模型的输入数据。矩阵中的一行对应科技报告的一篇摘要文档；列对应文档中的特征词；矩阵的行数即为文档数 M，列数即为第 m 篇文档中的特征词数，LDA 主题模型的输入数据结构如图 3 所示：

$$\begin{pmatrix} w_{11} & \cdots & w_{1N_1} \\ \vdots & \ddots & \vdots \\ w_{M1} & \cdots & w_{MN_M} \end{pmatrix}$$

图 3 LDA 输入数据结构

LDA 主题模型的核心是对参数 θ_m 和 φ_k 的估算，θ_m 和 φ_k 分别表示第 m 个文档中的主题概率分布和第 k 个主题中的特征词概率分布，两者分别是服从超参数 α 和 β 的狄利克雷的先验分布，其中 θ_m 和 φ_k 是初始自定义的。目前，对参数 和 有两种常用的算法：Gibbs 采样算法和 EM 算法。笔者使用 Gibbs 采样算法对参数进行估算，因为该算法对处理文中长文本有一定的优势，而且其空间复杂度和时间复杂度都较低。

Gibbs 采样算法参数估算过程如下：

（1）计算求得特征词－主题的联合概率分布：

$$p(w,z|\alpha,\beta)=p(w|z,w)p(z|\alpha) \quad 式(1)$$

（2）根据狄利克雷先验与贝叶斯法则，推算出狄利克雷分布期望为：

$$\theta_{m,k}=(n_m^{(k)}+\alpha_k)/(\sum_{k=1}^{K}n_m^{(K)}+\alpha_k) \quad 式(2)$$

$$\varphi_{k,t}=(n_k^{(k)}+\beta_1)/(\sum_{t=1}^{V}n_k^{(t)}+\beta_t) \quad 式(3)$$

式(2)和(3)中 $\theta_{m,k}$ 为在文档 m 中主题 k 的概率，$\varphi_{k,t}$ 为在主题 k 中特征词 t 的概率，$n_m^{(k)}$ 为在文档 m 中属于主题 k 的特征词的数量，$n_k^{(k)}$ 为特征词 t 属于主题 k 的次数。V 表示不同特征词的数量，K 为潜在主题数。

（3）通过 Gibbs 采样算法近似得出特征词－主题和主题－文档的联合概率分布：

$$p(z_i=k|z_{\neg i},w)\propto\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_{m,\neg i}^{(k)}+\alpha_k}\cdot\frac{n_{k,\neg i}^{(k)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t} \quad 式(4)$$

式(4)对应着一条从文档到主题到特征词的路径概率，其中路径的条数等于主题的个数 K，如图 4 所示：

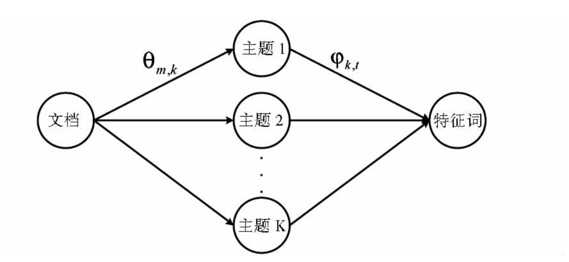


图 4 Gibbs 采样路径

Gibbs 采样就是在这 K 条路径中进行的，在推算出超参数 α 和 β 后验分布的同时，也得到了主题分布 θ_m 和特征词分布 φ_k 这两个参数。

LDA 主题模型处理后得到的输出结果为超参数 α 和 β 的后验估计和隐含参数 θ 和 φ。α 表示为在一篇科技报告文档摘要中，由所有未知主题生成概率所构成的 K 维向量（K 为主题数量）；β 表示科技报告文档摘要中每个特征词在每个主题中的生成概率，为一个 $V\times K$ 矩阵（V 为不同特征词数量）；θ 为一个表示每个主题在每个科技报告摘要文档中生成概率的 $M\times K$ 矩阵（M 为科技报告数量）；α 为一个表示指定科技报告摘要文档中每个特征词在每个主题中生成概率的 $V\times K$ 矩阵。

3.3 基于 Ward 与 K-means 的科技报告文档聚类模型设计

利用基于 Ward 与 K-means 的聚类算法对经过 LDA 模型处理后的科技报告文档进行再次聚类，以获得更准确有效的基于科技报告文档分类处理。

聚类算法具体流程如下：对上文经过主题模型处理后的每个科技报告摘要文档，用它所对应的主题概

chinaXiv:202308.00385v1

曲靖野，陈震，郑彦宁. 基于主题模型的科技报告文档聚类方法研究[J]. 图书情报工作,2018,62(4):113－120.

率向量 $\theta = [p_{z_1}, p_{z_2}, \cdots p_{z_n}]$ 来表示；用 Ward 算法对 LDA 主题模型挖掘出的带有主题概率分布的文档向量集进行第一次聚类；Ward 算法所确定类的数量 K_c 以及每个类中的均值作为 K-means 算法的两个初始条件；利用已确定初始条件的 K-means 算法对文档向量集进行第二次聚类，得到科技报告按照其所属主题概率分布的聚类结果。

3.4 评价方法

3.4.1 基于 LDA 模型的科技报告主题提取效果评价

笔者采用文献[20]中使用的查准率、查全率、值对 LDA 模型主题提取效果进行评价。查准率与查全率是判断数据挖掘结果优劣的两个重要度量指标，在信息科学领域被广泛应用。查准率是指挖掘出的正确信息占挖掘出的有效信息的比例，查全率是指挖掘出的正确信息占数据源中实际存在的正确信息的比例。两者公式表示如下：

$$P = N_c / N_p \qquad \text{式(5)}$$

$$R = N_c / N_R \qquad \text{式(6)}$$

在式(5)中，P 表示查准率，N_c 表示 LDA 模型提取的正确主题数，N_p 表示 LDA 模型提取的有效主题数；在式(6)中，R 表示查全率，N_R 表示文档集中实际存在的正确主题数。N_c、N_p、N_R 的值由项目组成员分两组独立设定，当两组结果不一致时再由相关领域专家以及判定。

查准率与查全率也存在局限性，两者有着相反的依赖关系，过高的查准率也会导致查全率降低，反之亦然，所以引入 F 值来调和查准率与查全率的对立关系，F 值的表达式如下：

$$F = 2PR/(P + R) \qquad \text{式(7)}$$

3.4.2 基于 Ward 与 K-means 的科技报告文本聚类效果评价 笔者采用聚类算法得出的总 F 值评价算法聚类效果[26]，总 F 值等于每个聚类 F 值的加权平均，如下所示：

$$F_{all} = \frac{\sum_{i=1}^{K_c} N_i F_i}{\sum_{i=1}^{K} N_i} \qquad \text{式(8)}$$

在式(8)中，F_{all} 为总 F 值，K_c 为聚类类数，i 为某个聚类，N_i 为聚类 i 中对象数量，F_i 为聚类 i 的 F 值，F_{all} 值越高说明算法的聚类效果越好。

4 实验流程与实验结果分析

4.1 实验流程

实验是在 Windows 7 系统环境下进行，计算机硬件 CPU 为 Intel i5 2.5GHz、内存 4G。中文分词、特征词提取及 LDA 建模分析采用的软件是 Python 3.6.0，文本聚类算法采用的软件是 Matlab 2012b。

笔者以国家科技报告服务系统 http://www.nstrs.cn/ 中 2013 年到 2017 年间与数字图像处理有关的科技报告为分析数据源，检索后经人工筛选到相关报告 1 842 篇。调用 Python 3.6.0 中的 Jieba 库对采集的中文摘要进行分词处理，选取词频在 95 到 40 之间的特征词(126 365 个)作为有效词构建特征词表。特征词表用 input.txt(1 842 行)来存储，并作为 LDA 模型的输入文件。在 Python 3.6.0 中使用主题模型工具包 Gensim 中的 LdaModel 函数来计算"文档－主题分布"和"主题－特征词"分布。在 Matlab 2012b 中，利用其自带的聚类函数工具包对 1 842 个包含主题信息的摘要文本向量进行聚类。

4.2 实验结果分析

笔者根据文献[15]设定超参数 α 和 β 的初始值分别为 0.01 和 1，主题数 K 的数值根据数据模型的困惑度确定为 42[27]。LDA 模型处理后得到特征词在 42 个潜在主题中的概率分布以及这 42 个潜在主题在 1 842 篇科技报告中文摘要中的概率分布。由于篇幅有限，笔者只列出其中 15 个主题中前 5 个特征词概率分布与其中 5 个文档中主题的概率分布，如图 5 和图 6 所示：

```
(0, '0.020*"分割" + 0.009*"识别" + 0.007*"区域" + 0.006*"检测" + 0.006*"边缘"
(1, '0.016*"鲁棒" + 0.012*"水印" + 0.012*"小波" + 0.012*"信息" + 0.009*"离散"
(2, '0.012*"测量" + 0.012*"距离" + 0.011*"遥感" + 0.009*"系统" + 0.008*"参数"
(3, '0.018*"检测" + 0.008*"精度" + 0.007*"提取" + 0.007*"差异" + 0.007*"边缘"
(4, '0.009*"提取" + 0.006*"特征" + 0.006*"向量" + 0.006*"信息" + 0.006*"检测"
(5, '0.012*"复原" + 0.006*"模型" + 0.006*"模糊" + 0.006*"退化" + 0.005*"物理"
(6, '0.008*"激光" + 0.007*"检测" + 0.007*"脉冲" + 0.006*"CCD" + 0.005*"去噪"
(7, '0.013*"重构" + 0.012*"小波" + 0.008*"系数" + 0.007*"分辨率" + 0.006*"规则"
(8, '0.011*"融合" + 0.006*"小波" + 0.006*"分辨率" + 0.005*"向量" + 0.005*"红外"
(9, '0.014*"智能" + 0.007*"识别" + 0.007*"人工" + 0.007*"人脸" + 0.006*"分割"
(10, '0.012*"目标" + 0.007*"追踪" + 0.007*"轨迹" + 0.006*"视觉" + 0.006*"LDA"
(11, '0.009*"三维" + 0.006*"测量" + 0.006*"重构" + 0.006*"建模" + 0.005*"立体"
(12, '0.013*"医学" + 0.013*"三维" + 0.012*"特征" + 0.012*"体层" + 0.009*"分割"
(13, '0.009*"压缩" + 0.007*"存储" + 0.006*"编码" + 0.006*"小波" + 0.006*"容量"
(14, '0.012*"配准" + 0.008*"模型" + 0.008*"分辨率" + 0.007*"像素" + 0.005*"信息"
```

图 5 15 个主题中前 5 个特征词概率分布

```
[(2, 0.30605443968829327), (14, 0.68238998519589866),]
[(11, 0.37219477937739726), (12, 0.61806736742224555)]
[(7, 0.98703701423316437)]
[(13, 0.98847735155914229)]
[(0, 0.52672407762458129), (3, 0.45994255135627699)]
```

图 6 5 个文档中主题的概率分布

从每个主题中的特征词可以推断出这 15 个主题对应着数字图像处理各个研究方向：图像分割、图像水印、图像测绘、目标检测、特征提取、图像复原、激光图像、图像重构、图像融合、人工智能、目标追踪、三维图像、医学图像、图像压缩、图像配准。主题之中有部分重复特征词，但是各个主题之间的边界基本清晰，在主

chinaXiv:202308.00385v1

ChinaXiv合作期刊

题标签的确定过程中，也咨询了相关领域的专家予以确认。

从 5 个文档的主题分布中可以得出，它们所包含的主题分别为：图像测绘和图像配准、三维图像和医学图像、图像重构、图像压缩、图像分割和图像检测。而这 5 篇科技报告的题目分别为：基于亮度序和图模型的多源遥感图像配准算法研究、现代医学成像与高维图像分析关键科学问题研究、基于结构先验约束的 PET 图像重建研究、基于 Grouplet 变换的 SAR 图像压缩感知编码、基于四元数的彩色图像边缘检测和分割方法研究。所以 LDA 模型得到的文档主题分布信息能够准确反映各个文档的研究内容。

为了证明 LDA 主题模型处理科技报告的有效性，笔者选取共词分析模型和 PLSA 模型与之进行对比。采用后两种模型分别对特征词表 input. txt（1 842 行）进行处理，三种对比模型所得到的查准率、查全率、值如表 2 所示：

表 2　不同模型对科技报告主题提取的查准率、查全率、*F* 值

对比模型	N_p	N_c	N_R	查准率	查全率	*F* 值
LDA	41	37	43	90.24%	86.05%	88.10%
共词分析	37	30	43	81.08%	69.77%	75.00%
PLSA	40	34	43	85.00%	79.07%	81.93%

从表 2 中的实验结果来看，LDA 模型主题提取的查准率、查全率、*F* 值都比较高，PLSA 模型次之，共词分析模型效果最差。可见主题模型相对于现有的主题识别方法，更加适合对科技报告文本的处理。

这主要是由科技报告文本自身的特点以及 LDA 模型本身的特性二者共同决定的。从科技报告数据对象本身特点来看：相对于科技期刊等其它科技文献，科技报告受篇幅限制较小，其对研究问题的阐述更加详细全面，话题内容范围广，相应导致基于其所提取的研究主题粒度更细，数量更多。从处理模型的角度来看：首先，共词分析方法只进行一次聚类运算来获得关于由关键词类簇所描述的主题，而主题模型多次迭代的运行过程会导致特征词和主题的概率在学习变化中趋于稳定，这种算法原理导致的差别会降低共词分析模型所提取主题中有效主题的数量（仅为 37）。其次，共词分析模型中一个特征词只能对应一个主题，而主题模型中特征词以不同概率分属于不同主题，这也导致共词分析模型所提取的正确主题数量较低（仅为 30）。以上可能是导致共词分析结果差于 LDA 模型和 PLSA 模型运行效果的主要原因。再次，相比于 PLSA 模型，LDA 模型中主题 – 特征词分布是可变的，而 PLSA 模型中主题 – 特征词分布是固定的，因此 LDA 的处理结果优于 PLSA。

根据上文所确定的 42 个主题数，将每篇科技报告摘要文档按照其最高的主题概率划分为 42 类，其分类所得结果如图 7 所示：

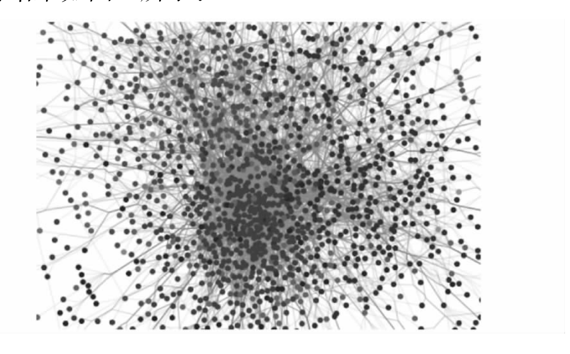


图 7　基于 LDA 主题模型的文档分类效果

在图 7 中，每个数据点表示一篇科技报告文档，相同颜色的数据点代表具有相同的最高概率主题的文档，这样的科技报告被认为是关于同一主题的文档，被划分为一类。由图 7 可见，同类别的科技报告数据点分散化严重，因此单纯按照 LDA 主题模型处理后的文档分类效果不理想。

使用笔者提出的 Ward 法与 K-means 法相结合的聚类算法对这 1842 个文档进行聚类，得到聚类效果图如图 8 所示：

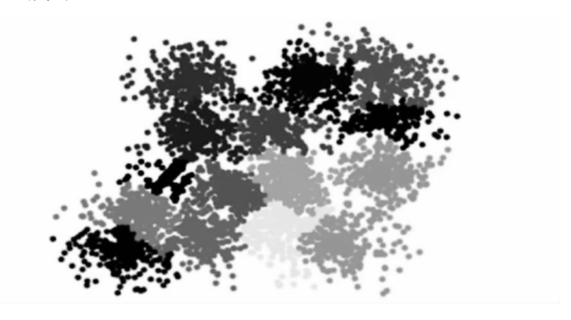


图 8　本文算法聚类效果

图 8 表明通过 Ward 法所确定的聚类类数为 15，比之前主题数 42 要少很多，这是因为这 42 个主题中有些主题研究方向相似，如运动视觉和目标追踪都属于机器视觉的研究方向。Ward 法最终将 1 842 个文档按照主题标签聚类成 15 个研究方向，K-means 法再将每篇科技报告文档聚类到最相近的研究方向中。从以上处理结果可见，笔者所设计的基于主题模型和聚类算法融合的科技报告文本分类方法，可以起到清晰有效划分文档挖掘主题的效果。

进一步从定量角度证明本文算法所实现聚类效果的准确性，把笔者提出的算法与单独使用 K-means 法

chinaXiv:202308.00385v1

和 Ward 法的聚类结果的总 F 值进行比较。K-means 法的 取值设定为 11 到 19,笔者提出的算法和 Ward 法的 取值由 Ward 法自动确定。三种算法的总 F 值如图 9 所示:

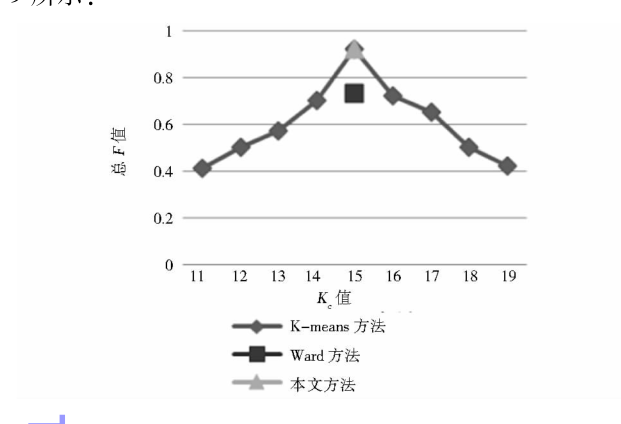


图 9　三种聚类算法结果总 F 值比较

图 9 表明当 K_c 为 15 时本文聚类算法与 K-means 聚类算法总 F 值最高并且相同,K_c 不等于 15 时 K-means 聚类算法总 F 值就会减少,表明其聚类效果降低,这进一步证明采用 Ward 法确定 K_c 可提高聚类效果,即本文聚类算法引入 Ward 法的合理性。当 K_c 为 15 时,Ward 法聚类总 F 值低于另外两种算法,这表明虽然 Ward 法确定的 K_c 是正确的,但是该方法没有再使用 K-means 聚类使每篇科技报告都正确聚类到每一类中,这导致了其聚类效果低于另外两种聚类算法在 K_c 为 15 时的聚类效果。因此,笔者提出的基于 Ward 与 K-means 的聚类算法是有效的。

5　结语

科技报告是国家科技实力与科技发展水平的重要体现,针对其关注度不充分的问题,笔者以国家科技报告服务平台为数据来源构建语料库,对其开展主题提取的实证研究,开辟了科技报告应用研究的新领域。基于 LDA 主题模型对科技报告文本进行主题挖掘,将生成的科技报告的文档－主题向量集作为聚类算法的输入,融合 Ward 与 K-means 算法对科技报告文档进行聚类,并通过实证分析证明了笔者所设计的科技报告文档聚类方法的有效性。笔者所设计的基于主题识别的文档聚类方法为科技报告的语义内容挖掘、知识服务的深度开展提供了方法支撑,虽然所采用的实验语料是基于科技报告文本所开展的,但本文的方法应用也可以为其他科技文献的文本挖掘提供一定的借鉴。

参考文献:

[1] VRETTAS G, SANDERSON M. Conferences versus journals in computer science[J]. Journal of the Association for Information Science & Technology, 2015, 66(12): 2674－2684.

[2] 侯人华,刘春燕,杜薇薇. 科技报告制度体系与形成模式研究[J]. 情报理论与实践,2014,37(1):51－54.

[3] 郭学武,朱江. 开放科技报告服务体系建设刍议[J]. 情报理论与实践,2011,34(9):82－84,126.

[4] 毛刚,贾志雷,侯人华. 情报学视角下的科技报告研究[J]. 情报杂志,2013 32(12):62－66,109.

[5] NTIS. The National Technical Information Service [EB/OL]. [2017－07－06]. https://www.ntis.gov.

[6] U.S. Department of Defense [EB/OL]. [2017－07－06]. https://www.defense.gov/.

[7] Office of Scientific and Technical Information. OSTI Databases [EB/OL]. [2017－07－06]. http://www.osti.gov.

[8] Science.gov [EB/OL]. [2017－07－06]. http://www.science.gov.

[9] OpenGrey. System for Information on Grey Literature in Europe. [EB/OL]. [2017－07－06]. http://www.opengrey.eu.

[10] 中国科学技术信息研究所. 国家科技报告服务系统[EB/OL]. [2017－07－06]. http://www.nstrs.cn.

[11] DUMAIS S, FURNAS G, LANDAUER T, et al. Using latent semantic analysis to improve access to textual information[C]//Proceedings of computer human interaction. Washington: Association for Computing Machinery, 1988:281－285.

[12] HOFMANN T. Probabilistic latent semantic indexing[C]//Proceedings of the 22th annual international SIGIR conference on research and development in information retrieval. Berkeley: Association for Computing Machinery, 1999:50－57.

[13] BLEI D, NG A, JORDAN M. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003,3 (3): 993－1022.

[14] TITOV I, MCDONALD R. Modeling online reviews with multi grain topic models[C]//Proceedings of 2008 WWW conference. New York: Association for Computing Machinery, 2008: 111－120.

[15] BLEI M. Probabilistic topic models [J]. Communications of the ACM, 2012, 55(4): 77－84.

[16] GRIFFITHS T, STEYVERS M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences, 2004, 101(S1): 5228－5235.

[17] 朱亮,孟宪学,赵瑞雪. 基于文献计量的科技监测方法与应用系统比较研究[J]. 数字图书馆论坛,2015,128(1): 52－56.

[18] 吕一博,康宇航. 基于共现分析的科技监测地图绘制及实证研究[J]. 科学学研究,2010,28(10): 1459－1466.

[19] 李湘东,张娇,袁满. 基于 LDA 模型的科技期刊主题演化研究[J]. 情报杂志,2014,33(7): 115－121.

[20] 关鹏,王曰芬,傅柱. 不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析[J]. 图书情报工作,2016,60(2):112－121.

[21] 王曰芬,傅柱,陈必坤. 采用 LDA 主题模型的国内知识流研究结构探讨:以学科分类主题抽取为视角[J]. 现代图书情报技

chinaXiv:202308.00085v1

ChinaXiv合作期刊

术,2016,37(4):8-19.
[22] 王平. 基于层次概率主题模型的科技文献主题发现及演化[J]. 图书情报工作,2014, 58(22): 70-77.
[23] 刘卫江. 基于主题模型的科技监测研究与实现[D]. 南京:南京理工大学,2014.
[24] SZEKELY G, RIZZO M. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method [J]. Journal of classification,2005,22(2):151-183.
[25] 奉国和,郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作,2011,54(2) : 41 - 45.
[26] 周昭涛. 文本聚类分析效果评价及文本表示研究[D]. 北京:中国科学院研究生院(计算技术研究所),2005.
[27] 关鹏,王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016,37(9): 42-50.

作者贡献说明:

曲靖野:提出研究思路和论文框架,撰写论文并进行研究内容的完善和修改;

陈震:收集实验数据,撰写论文并负责实验部分的设计仿真及论文修改;

郑彦宁:确定论文选题,提出修改意见。

Research on the Text Clustering Method of Science and Technology Reports Based on the Topic Model

Qu Jingye[1,2] Chen Zhen[1] Zheng Yanning[2]
[1] Information Technology and Media College of Beihua University, Jilin 132013
[2] Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] This paper explores the method of text clustering in the science and technology reports based on the topic model, develops new scientific literature technology monitoring areas, and puts forward a new semantic analysis method based on science and technology reports. [**Method/process**] Based on the national science and technology report service system, firstly, it conducted topic mining based on the LDA model after the text preprocessing; secondly, a clustering analysis based on the combination of K-means and Ward was carried out based on the text vector of the abstract containing theme distribution information. A proper text clustering method for the text mining suitable for the science and technical report was proposed. [**Result/conclusion**] The experimental results show that the LDA model can be effectively and accurately used in the topic mining of science and technology reports, and the clustering effect of the combination of Ward and K-means proposed in this paper is better than that of other traditional clustering algorithms in science and technology reports.

Keywords: science and technology report topic model LDA text clustering

chinaXiv:202308.00385v1

《知识管理论坛》被 DOAJ 收录

经国际知名开放获取平台 DOAJ(Directory of Open Access Journals)的评估,2017 年 2 月 10 日,《知识管理论坛》正式被其收录(查询地址: https://doaj.org/toc/2095-5472)。这对扩大本刊的传播范围,增加期刊对网络所有用户的内容可见度和使用率,提升期刊的学术影响力具有重要的意义。

DOAJ 是由瑞典隆德大学图书馆于 2003 年 5 月创建,以提供高质量开放获取期刊的查询和获取服务为目标。该平台收录的开放获取期刊都是经过同行评议或严格评审的学术性、研究性期刊,具有免费、全文、高质量的特点,对学术研究具有很高的参考价值。最初 DOAJ 仅收录 350 种期刊,截至 2017 年 2 月收录 9 200 多种开放获取期刊。

《知识管理论坛》(ISSN 2095-5472,CN11-6036/C)是由中国科学院主管、中国科学院文献情报中心主办、《图书情报工作》杂志社出版的纯网络(e-only)学术期刊,旨在推动知识时代知识的创造、组织和有效利用,促进知识管理研究成果的快速、广泛和有效传播。自 2013 年创刊以来,本刊坚持双盲的同行评议制度,对学术不端进行严格把控,遵循知识共享许可(CC)协议,实行立即、完全的开放获取出版,本次能顺利通过 DOAJ 的审核,是对本刊坚持高品质开放获取出版政策的认可,也必将推动本刊今后更快、更好地发展,推动全世界用户对本刊的利用,推动知识管理的研究与实践。

《知识管理论坛》编辑部